基于合成平均刺激的平均表征机制
——来自平均面孔吸引力的证据*

田欣然[1] 侯文霞[1] 欧玉晓[1] 易冰[1] 陈文锋[1*] 尚俊辰[2*]
([1]中国人民大学心理学系,北京 100872，wchen@ruc.edu.cn)
([2]辽宁师范大学心理学院,大连 116029，junchen_20081@163.com)

摘要：
[目的] 人类能够快速提取集合中的统计信息，形成平均表征。对于平均表征的产生机制，研究者提出整合集合成员以合成平均刺激，或计算集合成员特征值的平均值两种观点。以往研究中合成的平均刺激的特征值和计算成员特征值的平均值两种方式的结果相似，难以区分两种观点。由于多个面孔的吸引力评分的均值与用这些面孔合成的平均面孔的吸引力评分存在差异，本研究试图利用平均面孔吸引力的特性为平均表征的产生来源于合成平均刺激的观点提供了支持证据。
[方法] 利用多个面孔的吸引力平均值和合成平均面孔刺激的吸引力的差异，本研究使用经典的平均辨别任务（实验 1 和 2）和直接的吸引力评价任务（实验 3 和 4）为合成平均刺激的机制提供了支持证据。四个实验分别采用大容量面孔集合和小容量面孔集合探讨平均表征的形成，
[结果] 大、小集合都形成了平均刺激，并且平均表征的评定和加工结果与合成的平均刺激条件更接近；此外，集合吸引力出现了高评现象，但小集合高评现象更少出现。
[结论] 平均表征的产生可能来源于合成平均刺激的参与，而非简单的成员平均值计算；但平均刺激的作用受到集合大小的影响。

关键词： 平均表征；面孔吸引力；平均面孔
分类号： B842.2

Average percept in ensemble perception is based on morphed average object: Evidence from average facial attractiveness

TIAN Xinran[1], HOU Wenxia[1], OU Yuxiao[1], YI Bing[1], CEHN Wenfeng[1], SHANG Junchen[2],
[1](Department of Psychology, Renmin University of China, Beijing 100872, China)
[2](School of Psychology, Liaoning Normal University, Dalian 116029, China)

Abstract:
[Objective] Previous research demonstrated that ensemble perception of groups can be formed rapidly by extraction of the average of high-level complex features. However, it is unclear whether the average percept is the outcome of extraction from the characteristic value of the average stimulus (for example, average face) created from group members, or from calculation of the average value of group members' characteristic values. The above two values were confused with each other in prior research, since most average value of group members are similar as the characteristic

* 本文系中国人民大学科学研究基金（中央高校基本科研业务费专项资金资助，18XNLG10, 19XNLG20）、国家自然科学基金（31400869）、辽宁省社会科学规划基金（L19BSH005）项目成果之一；田欣然和侯文霞为共同第一作者

value of the average stimulus. However, the attractiveness rating of the average face created from a group of faces is usually systematically higher than the mean value of attractiveness ratings of this group of faces. Therefore, it is easier to explore how the ensemble coding of crowd face attractiveness (i.e. group attractiveness) is formed by comparing the attractiveness of the average face with the mean value of attractiveness rating of a group of faces. This could provide a useful approach to explore how the average percept is formed.

[Methods] The present study used the average discrimination paradigm (Experiment 1 & 2) and the scoring paradigm (Experiment 3 & 4) to clarify the mechanism of the formation of average percept by comparing the group attractiveness with the attractiveness of average face. To tackle this issue, whether the average face was presented in the group of faces or not was manipulated (conditions: Avg vs. NoAvg). Group size were also manipulated to explore whether group size modulated the formation of average percept. In the average discrimination paradigm, a group of faces served as group stimuli to be compare with the probe face for attractiveness. Participants were asked to judge which is more attractive between the group stimuli and the probe face. In the scoring paradigm, participants were asked to rate the attractiveness of group stimuli, the average face created from the group, and each face of the group in isolated manner. Each group consisted of twelve (in Experiments 1 and 3) or four faces (in Experiments 2 and 4). There were two kinds of groups: one is that all group members are original faces, without the average face. The other is that an average face morphed from other original faces was included in the group.

[Results] In Experiment 1, the proportions for judging probe average face more attractive than group attractiveness in the Avg condition was similar with the NoAvg condition. In Experiment 2, when the set size was four, the proportions for judging probe average face more attractive than group attractiveness were significantly higher in the NoAvg condition. Moreover, in Experiment 3, the ratings for group attractiveness were not significantly different between Avg and NoAvg conditions. This may indicate that the group attractiveness is based on the average face which was created from group members rather than the mean value calculated from group members' attractiveness. In addition, the diffusion model analysis showed that the coding time was longer for NoAvg condition, which indicated that the formation of average face needed cognitive resource. In Experiment 4, when the set size was four, the attractiveness rating of the average face was significantly higher than group ratings for the two kinds of groups. The different results in different group size may be interpreted as the outcome of weakened average percept caused by the salient individual face representations in small group. This was evident from several analyses: 1) group attractiveness and the attractiveness of morphed average face decreased with smaller set size (Experiment 4); 2) When the probe face was morphed average face, the proportion for judging probe face as more attractive than group attractiveness was greater, comparing with the condition when the probe was a new face whose attractiveness was similar with the morphed average face (Experiment 2); 3) The performance for the hypothesized condition with average percept included in the set is in between the conditions with/without real average face included

(Experiment 2-4). In addition, comparing with Experiment 1, the information accumulation speed in Experiment 2 is slower, the processing time of group attractiveness is longer, reflecting the disturbance of the individual face representation.
[Conclusions] Group attractiveness is based on the morphed average face, and the ensemble percept relies on the extraction from the average stimulus created from the group.
Keywords: average representation face attractiveness average face

1 引言

我们的视觉系统每时每刻接收到海量信息，这些信息很多是高度结构化的。这些结构化的信息彼此相似，以集合的形式存在。人们可以对这些集合进行知觉平均(perceptual averaging)，相当精确地抽取集合内所有成员的平均表征(average representation，Alvarez，2011; Haberman & Whitney，2012; Whitney & Yamanashi Leib，2018)，涉及大小、方向、明度、位置等低水平特征(Alvarez & Oliva，2008; Ariely，2001; Bauer，2009; Parkes et al.，2001)，也包括面孔身份、性别和表情等高级社会性信息(Haberman & Whitney，2007; Haberman et al.，2015; Li et al.，2016)。很多研究关注大脑中平均表征是如何产生的：是通过整合集合成员来形成一个平均刺激的表征还是通过个体成员特征的平均值(mean value)计算来完成平均任务(Maule & Franklin，2015; Whitney & Yamanashi Leib，2018)。以往研究中，平均表征通常是用集合成员的平均值来作为测量指标，隐含了平均表征等同于集合成员平均值的假设。然而，由于集合平均刺激的特征值和成员特征值的平均往往十分相似而难以区分，这种假设并不能作为证据来区分平均表征的产生是由于大脑中形成平均刺激的表征还是由于集合成员的平均值计算。因此，平均表征的形成机制仍然是一个有待解决的问题。解决这个问题的一个思路就是分离集合平均刺激的特征值和集合各成员特征值的平均，而平均面孔由于其吸引力通常都比合成平均面孔的成员吸引力平均值更高(Carragher et al.，2018; Komori et al.，2009)，很适合作为这个问题解决的切入点。为了有效区分平均刺激的表征和平均值计算的混淆问题，本研究通过利用集合面孔吸引力的平均表征和集合中所有面孔的吸引力平均值的差异性来考察知觉平均过程中是否形成了平均刺激的表征。

1.1 平均表征形成机制的争议

已往研究对平均表征的形成方式提出了两种主要解释，即基于分布式注意的整体编码和基于聚焦注意的个体编码。整体编码观点认为，视觉系统对集合刺激进行平行加工，因而被试能够准确表征集合平均值，却不能够对集合内的个体进行准确表征(Ariely，2001)。个体编码观点则认为，视觉系统将有限的注意资源集中在从集合中抽取的少数样本上，并对其进行精细加工，然后通过样本信息的平均值计算来推断集合的平均表征(de Fockert & Marchant，2008; Myczek & Simons，2008)。

从一般视觉加工的角度看，视觉信息的加工是分层级的。关于整体编码和个体编码的争论某种程度上可以归结为在平均表征形成过程中整体和个体的视觉加工层级优先性。最近，视觉加工的逆层级理论(Reverse Hierarchy Theory，Hochstein & Ahissar，2002; Hochstein et al.，2015)认为整体加工和个体加工存在着逆层级性，即统计表征作为一种由自下而上的快速过程构建的高水平表征优先于个体表征的觉察。逆层级理论认为，整体表征(如场景主旨gist)的意识知觉从高级皮层开始，是一种基于低级皮层输入的知觉过程；在视觉加工开始阶段，我们仅能有意识觉察到视觉场景的整体表征(如gist)，不能觉察到高水平整体表征的前因细节(antecedents，即构成整体表征的个体细节)；在这个优先的层级加工后，视觉系统才将注意导向到特定的低级皮层处理单元，提取局部细节信息，即在高级皮层的整体表征以自上而下的方式返回到局部加工(逆层级返回，reverse hierarchy return)以证实(或矫正)初步的整体表征估计值(Hochstein et al.，2015)。因此，根据逆层级理论可以推论，平均表征最先是大脑整合粗略的个体信息形成的，并非基于精确个体表征的平均计算；但在加工后期会受到个体表征的矫正。然而，这仅仅是个推论，有待更直接的证据证实。

1.2 集合吸引力高评现象和平均面孔吸引力

和表情、身份等其他面孔特征类似，学者也曾经推测面孔集合平均吸引力应等于每张面孔吸引力的平均值 (Abbas & Duchaine，2008; Brady & Alvarez，2015; Haberman & Whitney，2012)。早期研究发现，由三个不同吸引力水平的年轻男性面孔组成的集合的吸引力刚好等于三个男性的平均吸引力(Anderson，1965; Anderson et al.，1973)。

然而，研究者也发现了不同的结果。Van Osch 等人 (2015)系统地操纵了集

合的容量大小，发现当集合中的面孔数量超过 6 张，集合的吸引力评分要显著高于集合成员的评分的平均值。这被称为集合吸引力高评现象(group attractiveness effect)。小容量面孔集合在特定条件下也存在集合吸引力高评现象：例如 Willis (1960)发现在集合只有两张或三张面孔时，集合吸引力的评分要比成员的平均值更极端，高吸引力集合的评分会比成员平均评分更高。

Van Osch 等人(2015)认为集合吸引力高评现象可能的机制是面孔集合的知觉加工形成了平均面孔的表征，即面孔集合的平均吸引力表征并不是成员吸引力数值的平均，而是被试将集合中的所有面孔加以变形(morph)加工并融合成新的平均面孔(average-face)，从而影响集合面孔吸引力的评价。面孔吸引力与面孔平均性具有强相关(O'Toole et al.，1999; Rhodes et al.，2001)，即平均面孔的吸引力会由于平均表征自身携带的平均属性而得到提升，进而高于组成它的所有面孔的吸引力平均值 (Carragher et al.，2018)。

Van Osch 等人(2015)关于平均面孔的解释符合刺激集合形成平均刺激的表征的观点，但并没有解释集合吸引力平均值(Anderson et al.，1973; Luo & Zhou，2018)和集合吸引力高评不一致的矛盾。我们认为这并不矛盾，Anderson 等(1973)的研究都采取了较小的面孔集合(N=3 或 4)，而高评结果出现于较大的集合(N≥8)(Van Osch et al.，2015)。正是这种集合大小的差异可能引起不一致的结果：即相对于大集合面孔，集合较小时加工资源足以对个体成员精确加工，个体成员表征更突显，更容易干扰平均面孔表征。Li 等人(2016)的发现为这个观点提供了支持证据：在有限的加工资源下，平均表征相对于个体的表征具有优势性，个体表征的精确度较低；但如果加工资源比较充足，个体表征精确度上升，而平均表征的精确度则降低。根据逆层级理论，平均表征的形成不需要精确的个体表征，而数值的计算需要建立在较高的个体表征精确度之上。此外，平均面孔的吸引力也受集合面孔数量的影响，即小容量集合面孔合成的平均面孔吸引力也低于大集合合成的平均面孔(Langlois & Roggman，1990)。事实上，Van Osch 等(2015)也发现，当集合容量减少，高评现象出现的概率大幅下降。这可能是由于小容量集合形成的平均面孔吸引力相对不高或者平均面孔受到干扰，也可能是没有形成平均面孔而依赖于平均值计算。因此，小容量集合面孔吸引力的高评现象会减少，但其机制有待厘清。

1.3 问题提出和假设

目前集合平均表征的形成机制仍然停留在理论层面，尚未有直接的证据。合成平均刺激作为一种可能的机制，能够较好地解释集合表征加工相关的理论问题和实验现象。首先，为平均表征的整体编码和个体编码之争提供解决思路；其次，为视觉加工的逆层级理论提供实证支持；最后，为集合面孔吸引力的高评现象提供实证解释。本研究使用平均辨别任务(实验 1 和 2，mean discrimination paradigm，Haberman & Whitney，2009)和吸引力评价任务（实验 3 和 4），通过比较集合吸引力和平均面孔的吸引力探讨集合吸引力高评现象的机制，以进一步厘清知觉平均过程中是否形成了平均刺激的表征。平均辨别任务要求被试对单个刺激与集合平均表征进行知觉比较，用知觉比较后对集合平均表征的反应作为因变量来推断平均表征是否存在。吸引力评价任务要求被试对集合整体和平均刺激进行评价，直接反映平均表征的知觉过程。我们设置了集合中包含平均面孔的条件，如果集合加工中形成了平均面孔，那么集合原本有无平均面孔对结果将没有影响；如果没有形成平均面孔，那么集合包含平均面孔将促进平均辨别过程或是提升集合吸引力。实验 1 和 3 通过比较在大容量集合中包含或包含平均面孔刺激这两种条件下按键的比例来为集合平均面孔的形成提供更直接的证据，实验 2 和 4 通过比较不同集合容量中集合平均面孔与集合吸引力的关系是否发生变化来为平均表征的形成和平均值计算的争议提供实验数据支持。此外，通过扩散模型分析结果探究集合加工过程，为实验 1 和 2 提供信息加工过程方面的证据。根据平均表征形成机制的不同观点，可以对实验的结果有如下预测：

(1) 如果平均表征是通过集合成员的吸引力平均值计算，那么由于平均面孔的高吸引力，包含平均面孔刺激的集合会比不包含平均面孔刺激的集合吸引力更高，更接近平均面孔吸引力，从而当集合包含平均面孔，集合吸引力会提高（实验 3 和 4），在平均辨别任务中判断集合吸引力更高的倾向增加，判断探测刺激平均面孔吸引力更高的比例降低，并且不受集合大小的影响(实验 1 和 2)。

(2) 如果平均表征是通过形成平均刺激的表征来产生，那么集合中是否包含平均面孔刺激对平均辨别任务和评分任务没有影响(实验 1 和 3)，并且在平均面孔受干扰的小集合中，包含平均面孔刺激的集合吸引力更接近平均面孔吸引力（实验 4），从而判断探测刺激平均面孔吸引力更高的比例降低(实验 2)。

(3) 在假设 2 的基础上，如果小集合平均面孔受干扰是高评现象减少的原因，那么在个体表征较为突出的情况下，集合吸引力和平均面孔吸引力的差异在不包含平均面孔刺激的集合条件下更大（实验 4），进而，小集合判断探测刺激平均面孔吸引力更高的比例在不包含平均面孔刺激的集合条件下更高(实验 2)。

（4）如果小集合平均面孔吸引力相对大集合而言较低是高评现象减少的原因，那么小集合中平均面孔吸引力下降导致集合吸引力和平均面孔的差异更小（实验 3 和实验 4 对比），进而，小集合判断探测刺激平均面孔吸引力更高的比例更低(实验 1 和实验 2 对比)。

2 实验1：大容量面孔集合加工中是否形成平均面孔

实验 1 通过平均辨别任务，要求被试选择集合吸引力和平均面孔吸引力之间较高的一个，操纵集合中是否出现平均面孔，进而判断平均面孔是否形成。

2.1 方法

（1）被试

采用 GPower 以统计功效 power=0.8，中等效应量 f=0.25 和重复测量 2(自变量集合类型：2 个水平)为参数估计的最小样本量为 N=34。实际招募中国人民大学在校生 34 名(其中 18 名女性)，平均年龄 20.75 岁，标准差 2.02，右利手，视力或矫正视力正常。本研究中所有被试均签署知情同意书，实验得到了中国人民大学心理学系伦理委员会的批准。

（2）实验材料

为了产生足够的组间差异，选取了互联网材料作为吸引力极高和极低的面孔图片，去除头发、脖子和耳朵，将面部轮廓剪成椭圆形，并转换成灰度图像来进行标准化。部分面孔材料选自中国化面孔情绪图片系统（王妍, 罗跃嘉，2005）中的女性–中性情绪库。中国化面孔情绪图片系统中选取的图片与互联网材料共同组成原始材料。所有材料都经过 20 名中国大学生在吸引力水平和情绪效价（均为 101 点量表评分）上的评分(10 名女性，平均年龄为 20.54 岁，标准差 2.17)，其中被评价为非中性（与评分 50 有显著差异）的材料被剔除。选定的材料效价评分（M = 49.94，SD = 0.77）与中性（评分 50）没有显著差异，$t(19)$ = 0.35，p = 0.732。评定后选择的原始面孔 30 张，其中包含 6 张互联网材料（其中 4 张属于高吸引力组，2 张低吸引力组），其余 24 张图片从中国化面孔情绪图片库中选

取。

根据吸引力评分将原始材料分为吸引力高中低三组，每组 10 张图片。高、中、低吸引力组的平均分分别为 76.33; 43.12; 22.42。重复测量方差分析检验三组评分之间的差异发现差异显著，$F(2,38) = 148.64$，$p < 0.001$，$\eta_p^2 = 0.89$。两两比较差异显著，低吸引力组与中吸引力组：$t(19) = 7.91$，$p < 0.001$，Cohen's $d = 3.63$；中吸引力与高吸引力组：$t(19) = 9.77$，$p < 0.001$，Cohen's $d = 4.48$；低吸引力组与高吸引力组：$t(19) = 15.92$，$p < 0.001$，Cohen's $d = 7.30$。随后我们设置来自不同吸引力分组的面孔组成的集合，每个集合中的图片数量包含 11 张，12 张两种情况，实验 1 共 60 个集合。在单个集合中不同吸引力分组的图片被使用的次数不同，而每个实验所用的所有集合中，不同吸引力分组的图片总共被使用的次数是相同的。

随后，用面孔合成软件 Abrosoft FantaMorph（Abrosoft Fantamorph.5.4.8，www.fantamorph.com）将不同面孔集合的平均面孔制作出来，共有 365 张。该软件可以将两张面孔按照一定的比例融合，将面部各特征以众多关键点来标注，如嘴角的位置，大小，弧度，随后取关键点的平均值来合成图像。例如，当我们希望制作 4 张原始面孔的平均面孔，就将原始面孔两两一组，再按照 50:50 的比例进行合成取中，将合成的两张图片再次按照 50:50 比例合成，就相当于每张原始面孔在合成面孔的贡献比例为 25%，得到了 4 张原始平均面孔的平均面孔。如果要制作 3 张原始面孔的平均面孔，则控制每张面孔的贡献比例为 33.3%即可。

所有的面孔图片再次经过 20 名中国人民大学在校生的吸引力评定（10 名女性，平均年龄为 20.35 岁，标准差 2.03），作为事先评定的得分。

所有实验材料使用 24 英寸的 Dell 显示屏呈现，分辨率为 1920×1080，灰色背景，被试直坐时双眼距离显示屏距离约为 70cm。

（3）实验设计

采用单因素被试内设计，自变量为集合类型(无平均面孔的集合 G1 vs. 有平均面孔的集合 G2)，因变量为判断平均面孔吸引力更高的比例和扩散模型分析得到的反应决策指标（信息累积速率 v、阈限差值 a 和非决策加工时长 t0，详见结果部分）。

（4）实验程序

实验采用平均辨别任务，先呈现集合刺激，再呈现探测刺激。探测刺激为集合平均面孔、集合成员面孔、非成员非平均面孔。由于和平均面孔以及集合平均值的大小关系不确定，难以对结果进行推断，后两种刺激类型在本研究中只作为反应填充刺激（控制条件）。

在集合刺激类型上，使用 12 张原始面孔组成集合，即是“不包含平均面孔的集合 G1”水平，如果使用 11 张原始面孔组成集合，并将集合成员的平均面孔作为新成员进入集合中，即是“包含平均面孔的集合 G2”水平。包含平均面孔的集合中，平均面孔的位置随机呈现。在探测刺激类型上，呈现集合成员的平均面孔即是“集合平均面孔”水平，当集合刺激包含平均面孔时，相当于平均面孔出现两次；“呈现集合成员之一”的水平中，呈现集合中除了平均面孔以外的其他成员面孔之一；呈现集合刺激中没有出现过的面孔即是“新面孔”水平。

在每个试次中，被试首先注视中心点 1000ms，随后，他们看到呈现在屏幕上的集合刺激 2000ms，之后呈现空屏 500ms，随后呈现一张探测面孔，呈现到出现反应为止。要求被试按 F 或 J 键判断集合刺激的整体吸引力和探测刺激的个体吸引力哪个更高，共 180 个试次，各个条件混合随机呈现，每 60 个试次休息一次。在探测面孔为新面孔和集合成员之一两种条件下，一半探测刺激在预评中的吸引力高于集合刺激成员吸引力平均值，一半低于平均值(如图 1)。

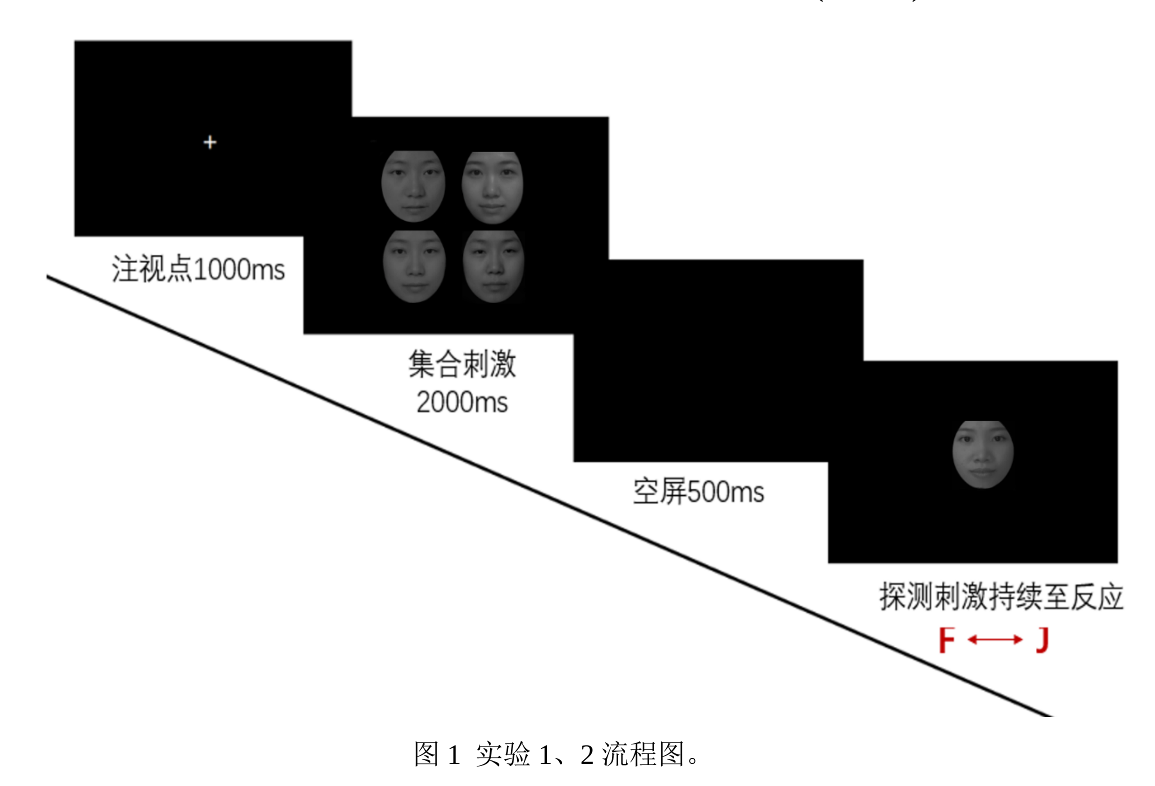


图 1 实验 1、2 流程图。

集合刺激以4×3矩阵呈现，单张面孔图片的视角为5.69° ×6.53° 。探测刺激材料是一张单独的面孔，呈现在屏幕中央，图片尺寸与集合成员刺激尺寸一致。

（5）扩散模型

根据逆层级理论，平均面孔的形成尽管快速，但仍然需要信息累积，因而有可能在包含和不包含平均面孔的集合之间产生决策反应差异。为了考察这种可能的差异，我们分析了反应决策信息，采用扩散模型(the diffusion model；Ratcliff，1978; Ratcliff & McKoon，2008) 将不同的认知过程进行分解，综合利用反应时分布与反应准确性结果，进一步分析包含和不包含平均面孔对集合面孔吸引力的知觉机制。该模型可以将分解的认知过程对应到不同的模型参数中(Voss et al.，2013)。

扩散模型的基本假设是：在快速的二选一任务中，信息从起始点不断累积直到达到某反应的阈限标准后激活反应。基本扩散模型(Ratcliff，1978)有四个参数(如图 2)，分别为：

1) 漂移率(drift rate)，记为 v，表明信息累积的速率；

2) 阈限差值(threshold separation)，记为 a，表明做决策所需要的信息量；

3) 起始点(starting point)，记为 z，表明决策前的预先偏向；

4) 非决策加工时长(duration of nondecisional processes)，记为 t0，包含编码、反应执行等非决策的时间。

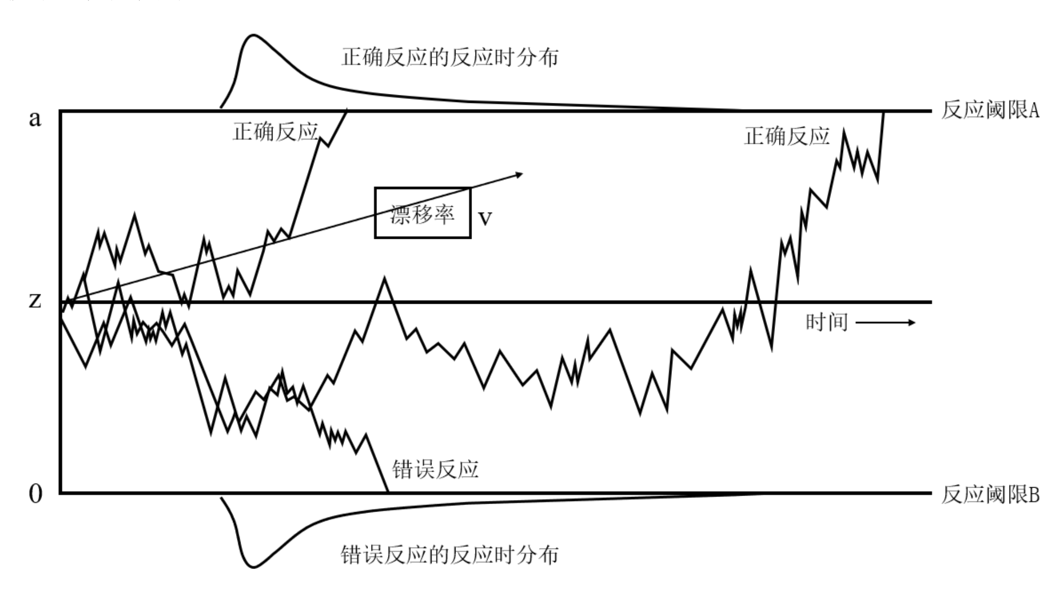


图 2　扩散模型(翻译自 Ratcliff & McKoon，2008，Figure 2)。图中展示了扩散模型的三条路径样例。信息从起始点(z)以平均速率(v)开始逐渐累积，直到达到反应 A 的阈限(a)或反

应 B 的阈限(0)。由于随机噪音，这些路径在每个试次之间都有所变异。

2.2 结果

（1）按键反应结果

我们根据预评吸引力分数计算了在探测刺激类型为控制条件刺激的反应正确率，以此确定被试进行了充分理解和正确反应，同时验证事先评定的吸引力评分是否适用于本实验的被试。根据预评分数计算集合成员的吸引力平均值，再和预评的探测面孔吸引力比较来确定正确反应，结果表明探测刺激类型为新面孔和集合成员之一两种条件下的总正确率达到 84.72%，远高于随机反应，$t(33) = 28.21$，$p < 0.001$，95%CI = [0.32，0.37]，Cohen's $d = 9.82$。说明被试的吸引力判断和事先评定基本一致。根据事先评定的得分，我们分别统计了实验 1 中不含平均面孔的集合中探测刺激为平均面孔的条件下所有集合成员吸引力的平均值 $M1 = 49.19$，同时假设该集合合成了平均面孔并计算包含了该平均面孔的集合成员吸引力平均值，也就是假设生成了平均面孔并将其吸引力计算进集合的成员平均值 $M2 = 50.49$。$M1$ 和 $M2$ 的差异表明，合成平均面孔提高了集合吸引力平均值，$t(19) = 22.82$，$p < 0.001$，95%CI = [1.14，1.37]，Cohen's $d = 10.47$。

我们统计了实际反应中判断探测刺激平均面孔吸引力更高的比例。无论集合中包含和不包含平均面孔，被试判断平均面孔吸引力更高的比例（不包含平均面孔 G1：84.03%，包含平均面孔 G2：83.55%）都显著高于随机概率（50%），$t(33) = 8.16$，$p < 0.001$，95%CI = [0.25，0.42]，Cohen's $d = 2.84$；$t(33) = 10.31$，$p < 0.001$，95%CI = [0.27，0.40]，Cohen's $d = 3.59$。当探测刺激为平均面孔，判断平均面孔吸引力更高的比例在包含平均面孔、不包含平均面孔的集合类型之间没有显著差异(如图 3)，$t(33) = 0.11$，$p = 0.912$，95%CI = [–0.10，0.11]，表明有无平均面孔对集合吸引力的知觉辨别没有显著影响。

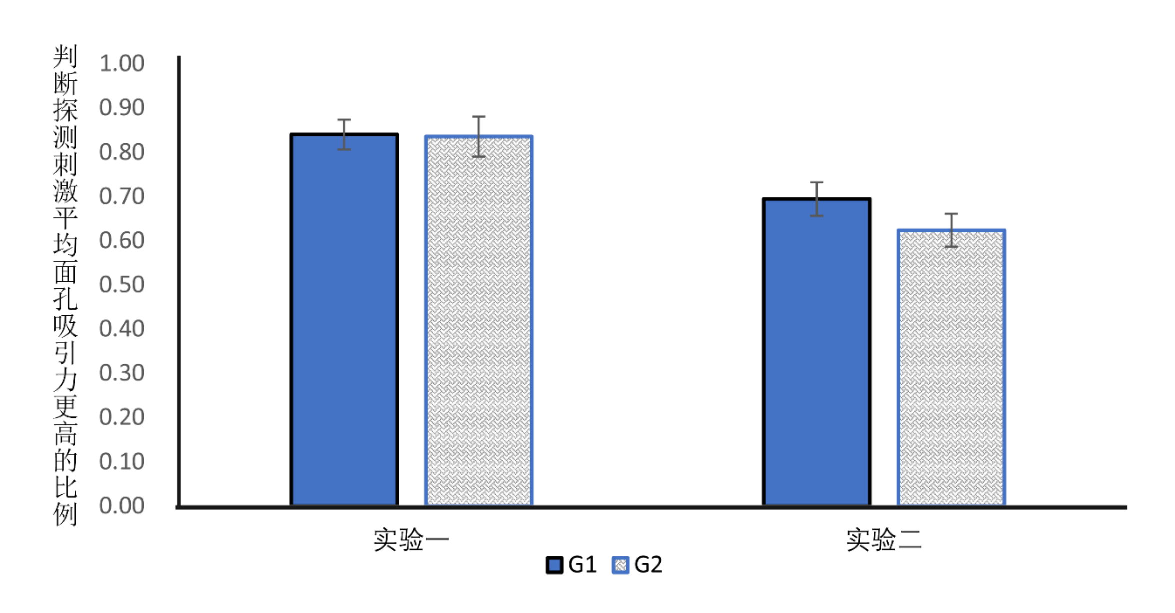


图3　不同条件下被试判断平均面孔吸引力更高的比例（注：G1为不包含平均面孔的集合、G2为包含平均面孔的集合）

（2）扩散模型分析

前述比例分析虽然表明有无平均面孔对集合吸引力的知觉辨别没有显著影响，但并不清楚对辨别决策过程(如辨别时间、决策标准等)有无影响。这里我们采用层级扩散模型方法(the hierarchical diffusion model，HDM; Vandekerckhove et al.，2011)进行数据的模型拟合。HDM 分析的优势在于在模型参数计算时考虑被试之间的个体差异。此外模型上下限分别设定为正确反应和错误反应，在探测刺激为平均面孔条件时，将判断探测刺激吸引力更高设定为正确反应。由于对于正确和错误反应不存在预先的反应偏向，因此模型将起始点(z)设置为 a/2。模型的其他参数设定为随研究变量(集合类型，探测刺激类型)变化。通过层级扩散模型拟合，最终得到每个被试在每个条件下的漂移率 v，阈限差值 a 和非决策加工时长 t0，并进行统计检验分析(如图 4)。

扩散模型是对每个被试进行单独拟合，一般认为，如果模型拟合优度参数 Rhat 小于 1.05(Vehtari et al., 2019)，则拟合度较优，我们对实验 1 的拟合结果进行单样本 t 检验发现，各拟合参数（M = 1.00）均显著小于 1.05，表明模型拟合良好。

以模型参数为因变量的t检验表明，在信息累积漂移率v和阈限差值a上，集合是否包含平均面孔没有显著差异，$t(33) = 0.48$，$p = 0.632$；$t(33) = 1.72$，$p = 0.096$。表明其不影响对集合吸引力和平均面孔的辨别；但非决策加工时长t0受到

了集合是否包含平均面孔条件的影响，集合包含平均面孔条件所需的t0更短，$t(33) = 2.57$，$p = 0.015$，95%CI = [0.01，0.06]，Cohen's $d = 0.90$。

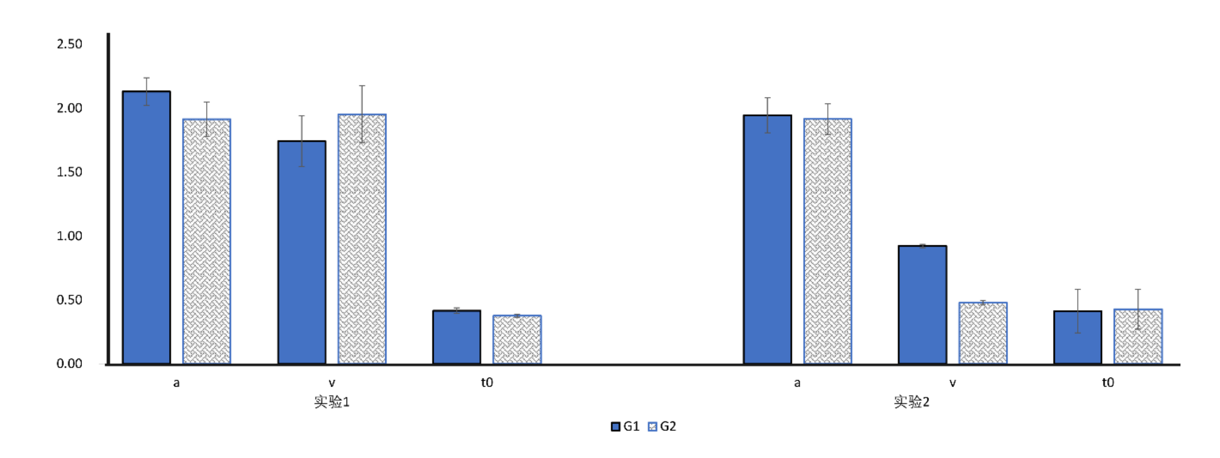


图4 实验1、实验2获得的层级扩散模型拟合结果（注：G1为不包含平均面孔的集合、G2为包含平均面孔的集合）

2.3 讨论

实验 1 的结果表明，容量为 12 的集合面孔吸引力确实存在高评现象。这种高评现象来自合成平均面孔的更高吸引力，并影响了被试对集合吸引力和探测刺激平均面孔的比较，使得与实际包含平均面孔的集合条件结果没有差异。实验结果模式符合假设 2 的合成平均刺激，不支持假设 1 的平均值计算。因此，平均辨别并不是简单地通过计算集合原有成员的平均值，而是形成了集合平均面孔。由于平均面孔的高吸引力，有平均面孔集合所有成员的平均值高于无平均面孔的集合的成员平均值，更接近探测刺激平均面孔，从而探测刺激和集合吸引力更不易区分。因而，如果是根据平均值计算来完成平均辨别任务(即成员平均值是比较标准)，故而有平均面孔条件下判断平均面孔吸引力更高的比例将低于无平均面孔条件。但是，结果与此预测相反，判断平均面孔吸引力更高的比例在包含平均面孔、不包含平均面孔的集合类型之间没有显著差异。因此，更可能的是，人们在知觉集合刺激时将成员合成为一张新面孔（即具有高吸引力的平均面孔），从而对不包含平均面孔的集合吸引力造成显著的提高；而在包含平均面孔的条件中，整个集合形成的平均面孔应该等同于其中的 11 张原始面孔形成的平均面孔，也就是集合中已经出现的这张平均面孔。那么平均面孔出现在集合中就不应当对集合吸引力造成显著的提高；故包含/不包含平均面孔的条件中选择集合整体吸引力更高的比例不存在差异。

扩散模型分析结果表明，在没有实际平均面孔的输入时，编码加工等非决策时间更长，这说明平均辨别任务过程中形成平均面孔需要加工时间和资源的投入。这个过程是很快的（约 400ms），因而对决策影响不大，从而决策信息累积速度与有实际平均面孔输入时没有显著差异。

总结而言，实验 1 的结果表明集合吸引力的判断过程形成了平均面孔，从而导致集合吸引力高评。那么，有什么因素会影响平均面孔的形成呢？近期研究发现，高级特征如面孔表情的平均表征是个容量有限的过程(Ji et al.，2018)，受到加工资源的制约(Li et al.，2016)。那么，面孔吸引力的平均表征是否也受到容量的影响？从以往的结果看，答案似乎是肯定的。例如，Van Osch(2015)发现，在集合容量为 4~6 的集合中，集合吸引力高评现象出现的可能性很低。以往发现面孔集合吸引力相当于成员平均值的研究也都采用的是较小的集合(Anderson，1965; Anderson et al.，1973)。这是否说明小容量集合中没有平均面孔形成，小集合面孔吸引力判断存在与大容量集合不同的机制？但还存在另外一种可能，即平均面孔确实形成了但受到干扰。为了分离这两种可能，实验 2 考察了小容量集合吸引力与平均面孔的关系。

3 实验2：集合容量对面孔集合加工的影响

实验 2 采用 4 张面孔组成的集合，采用与实验 1 相同的实验设计和程序，考察了容量对集合吸引力与平均面孔的关系的影响。

3.1 方法

（1）被试

采用 GPower 以统计功效 power=0.8，中等效应量 f=0.25 和重复测量 2(自变量集合类型：无平均面孔的集合 G1 vs. 有平均面孔的集合 G2)为参数估计的最小样本量为 N=34。实际招募中国人民大学在校生 35 名大学生，排除一名记错按键方向的被试，有效被试 34 名(17 名女性)，平均年龄 20.68 岁，标准差 2.27，右利手，视力或矫正视力正常。

（2）实验材料

与实验 1 相同，但集合刺激只包含 4 张图片。在集合刺激类型上，使用 4 张原始面孔组成集合，即是“不包含平均面孔的集合 G1”水平，如果使用 3 张原

始面孔组成集合，并将集合成员的平均面孔作为新成员进入集合中，即是“包含平均面孔的集合 G2”水平。探测刺激类型包括集合平均面孔、集合成员面孔、非成员非平均面孔，探测刺激的后两种刺激类型是控制条件刺激。

集合刺激以 2×2 矩阵呈现，图片的尺寸为 8.19°× 9.43°。探测刺激材料是一张单独的面孔，呈现在屏幕中央，图片尺寸与集合刺激尺寸一致。

（3）实验设计

和实验 1 相同。每个被试在主任务完成后还对每张图片进行了吸引力评分。

（4）实验程序

与实验 1 相同。

3.2 结果

（1）按键反应结果

根据事先评定的得分，在探测刺激类型为新面孔和集合成员两种条件下的反应正确率为 84.17%，显著高于随机水平，$t(33) = 16.84$，$p < 0.001$，95%CI = [0.31，0.39]，Cohen’s $d = 5.83$，以此确定被试确实充分理解和正确反应。与实验 1 相似，分别统计了探测刺激为平均面孔的条件下不含平均面孔的集合中所有成员吸引力的平均值 $M1 = 47.82$，同时也假设该集合生成了平均面孔并计算包含了该平均面孔的成员均值 $M2 = 49.73$。$M1$ 和 $M2$ 的差异表明，平均面孔也同样提高了小容量集合吸引力平均值，$t(29) = 6.68$，$p < 0.001$，95%CI = [1.44，2.47]，Cohen’s $d = 2.48$。根据被试主任务后的评定，发现平均面孔的吸引力（$M = 55.18$, $SD = 11.02$）高于集合成员面孔吸引力的平均值（$M = 51.71$, $SD = 11.76$），$t(33) = 2.35$，$p = 0.020$，95%CI = [0.51，7.05]，Cohen’s $d = 0.820$。

对实验 2 和实验 1 中平均面孔为探测刺激条件下的平均面孔吸引力进行比较，结果表明小集合面孔形成的平均面孔吸引力更低，57.20 vs. 65.61，校正 $t(41.7) =100.61$，$p < 0.001$， 95%CI = [8.26，8.60]，Cohen’s $d = 24.53$。对实验 2 和实验 1 中平均面孔(探测刺激)和集合平均值的差值(9.51 vs. 16.43)进行跨实验比较，发现实验 2 小集合平均面孔和集合平均值的差异更小，校正 $t(53.8) = 112.13$，$p < 0.001$，95%CI = [6.70，6.94]，Cohen’s $d = 27.53$。因而，小集合中判断探测刺激平均面孔吸引力更高的比例将下降。结果确实如此，实验 2 中判断平均面孔吸引力更高的比例显著低于实验 1(66.57% vs. 83.79%)，校正 $t(63) = 3.37$，$p = 0.001$，

95%CI = [0.07，0.27]，Cohen’s d = 0.85。

统计检验结果表明，被试倾向于认为平均面孔的吸引力要高于集合的吸引力。判断探测刺激平均面孔吸引力更高的比例显著高于随机概率，$t(33) = 4.60$，$p < 0.001$，95%CI = [0.09，0.24]，Cohen’s d = 1.60。而且对于判断平均面孔吸引力更高的比例，集合中不包含平均面孔的条件显著高于包含平均面孔的条件，$t(33) = 3.77$，$p = 0.001$，95%CI = [0.03，0.12]，Cohen’s d = 1.31(如图 3)。

由此可见，小集合的加工结果确实与大集合不同，为了探究这种差异的存在是由于加工机制不同还是平均面孔受到干扰，我们对不含平均面孔的集合中是否形成了平均面孔进行了检验。我们通过匹配选择出与平均面孔吸引力接近的新面孔探测刺激，两类面孔的吸引力均值分别为 54.80 vs 54.11，$t(46) = 0.18$，$p = 0.859$，但探测刺激为平均面孔时探测刺激吸引力被判断为更高的比例(69.12%)仍然高于新面孔条件(52.01%)，$t(33) = 4.84$，$p < 0.001$，95%CI = [10.21%，24.88%]，Cohen’s d = 1.69。即使进一步匹配选择出比平均面孔吸引力更高的新面孔刺激(72.33 vs. 60.86，$t(42) = 3.85$，$p < 0.001$，95%CI = [5.54%，17.49%]，Cohen’s d = 1.19)，探测刺激为平均面孔时探测刺激吸引力被判断为更高的比例(71.01%)仍然高于新面孔条件(61.31%)，$t(33) = 2.62$，$p = 0.013$，95%CI = [2.24%，17.13%]，Cohen’s d = 0.91。这说明在辨别过程中平均面孔并不是作为新面孔出现的，而更可能是集合呈现时形成了平均面孔。

我们还计算了以实际集合成员平均值 *M*1 和假设形成了平均面孔后的集合成员平均值 *M*2 作为比较标准时的正确率 *Acc*1 和 *Acc*2。由于平均面孔评分和集合平均值的差值较大(平均差异 10.40)，容易出现天花板效应，我们只选择了差值低于 10 的试次计算 *Acc*1 和 *Acc*2。如果形成了平均面孔，那么集合吸引力提高，和探测刺激的平均面孔的差异减小，因而正确率将降低，*Acc*2 低于实际平均值的正确率 *Acc*1，从而接近实际包含平均面孔的集合 *Acc*0。t 检验结果表明，*Acc*2(50.34%)和 *Acc*1(53.03%)，$t(33) = 1.18$，$p = 0.249$，Cohen’s d = 0.42，Acc2 和 *Acc*0(45.79%)，$t(33) = 1.46$，$p = 0.154$，Cohen’s d = 0.51，都没有显著差异。以三个正确率为自变量的三个水平进行单因素趋势分析(trend analysis)，结果表明从 *Acc*1 到 *Acc*2 再到 *Acc*0 存在线性递减趋势，$F(1，33) = 4.21$，$p = 0.048$，$\eta_p^2 = 0.11$。这些结果表明，尽管不含平均面孔的小集合检测到的平均面孔效应无法

完全与包含平均面孔的条件相等，但合成平均面孔依然在一定程度上起作用。

（2）扩散模型分析

实验 2 同样通过层级扩散模型拟合，最终得到每个被试在每个条件下的漂移率 v，阈限差值 a 和非决策加工时长 t0，并进行统计检验分析。各拟合参数（*M* = 1.00）均显著小于 1.05，表明模型拟合良好。

以模型参数为因变量的 *t* 检验表明，在阈限差值 a 和非决策加工时长 t0 上，集合是否包含平均面孔对集合吸引力和平均面孔的辨别没有影响，$t(32) = –0.63$，$p = 0.533$，$t(32) = 0.72$，$p = 0.095$；但信息累积漂移率 v 受到了集合是否包含平均面孔条件的影响，集合不包含平均面孔条件的信息累积更慢，$t(33) = –4.775$，$p < 0.001$，95%CI = [–0.63，–0.25]，Cohen's *d* = 1.66(如图 4)。

层级扩散模型的统计结果在实验间存在差异(如图 4)。在非决策加工时长 t0 上，集合包含平均面孔与否和集合容量存在交互作用，$F(1,63) = 14.03$，$p < 0.001$，$\eta_p^2 = 0.18$。简单效应分析发现，在集合包含平均面孔条件下，实验 1 的 t0 显著小于实验 2 的 t0，$t(63) = –2.568$，$p = 0.013$，95%CI = [–0.09，–0.01]，Cohen's *d* = 0.65。在漂移率 v 上，同样存在集合包含平均面孔与否和集合容量的交互作用，$F(1,63) = 9.63$，$p = 0.003$，$\eta_p^2 = 0.13$，简单效应分析发现，无论集合包含或不包含平均面孔，实验 1 的 v 显著大于实验 2 的 v，$t(63) = 5.51$，$p < 0.001$，95%CI = [0.94，2.00]，Cohen's *d* = 1.39；$t(63) = 3.16$，$p = 0.002$，95%CI = [0.30，1.34]，Cohen's *d* = 0.80。

3.3 讨论

实验 2 发现，在不包含平均面孔的条件中，判断平均面孔吸引力更高的比例显著高于包含平均面孔的条件，说明平均面孔出现在集合中，显著地增加了集合的整体吸引力。由此可见，当集合容量为 4，被试主观形成的平均面孔表征被抑制或者没有形成。如果是平均面孔表征没有形成，则平均面孔作为探测刺激应该和新面孔类似，但分析表明平均面孔和新面孔的探测结果完全不同。因而，小集合面孔也形成了平均面孔。这些结果模式符合合成平均刺激的假设 2，不支持平均值计算的假设 1。

尽管小集合面孔形成了平均面孔，但平均辨别的反应模式和实验 1 大集合并不相同。这可能是因为小集合中形成的平均面孔更容易受干扰，因而包含平均面

孔刺激的集合吸引力更接近平均面孔，从而判断平均面孔吸引力更高的比例降低。这也反映在被试的反应决策参数上，决策所需的信息累积速度相对实验 1 更慢，所需的加工时间更长。扩散模型的结果表明，在有实际平均面孔的输入时，辨别决策将更容易，表现为决策信息累积更快。这些结果说明平均辨别任务过程中形成了平均面孔，只是受到了干扰。这种干扰还体现在小集合判断平均面孔吸引力更高的比例比实验 1 更低。实验结果表明，这可能来自两个原因：一是小集合平均面孔受干扰（假设 3），那么集合吸引力和平均面孔的差异在不包含平均面孔刺激的集合条件下更大，从而小集合判断平均面孔吸引力更高的比例在不包含平均面孔刺激的集合条件下更高；二是小集合平均面孔吸引力更低，集合和平均面孔的差异更小（假设 4）。

实验 1 和实验 2 的结果来自于相对间接的平均辨别任务。为了提供更直接的证据，实验 3 和实验 4 采用评分任务对实验 1 和实验 2 的结果进行进一步验证。

4 实验3：大容量面孔集合的评分

实验3采用大容量面孔集合进行评分任务。为不同容量下集合吸引力和平均吸引力的关系提供更直观的证据。

4.1 方法

（1）被试

采用 GPower 以统计功效 power=0.8，中等效应量 f=0.25 和单因素 5 水平(评分类型：不包含平均面孔的集合的成员均值 M1、不包含物理平均面孔集合但将平均面孔计算在内的成员均值 M2、不包含平均面孔的集合 G1、包含平均面孔的集合 G2、平均面孔 Avg)为参数估计的最小样本量为 N=21。实际招募中国人民大学在校生 29 名大学生，有效被试 29 名(15 名女性)，平均年龄 22.14 岁，标准差 3.17，右利手，视力或矫正视力正常。

（2）实验材料

实验刺激与实验 1 相同。

集合刺激包含 12 张面孔，集合刺激以 4×3 矩阵呈现，单张面孔图片的视角为 5.69°×6.53°。

在评分类型上，使用 12 张原始面孔组成集合，即是“不包含平均面孔集合”

水平；如果使用 11 张原始面孔组成集合，并将集合成员的平均面孔作为新成员进入集合中，即是“包含平均面孔集合”水平，将集合成员面孔和平均面孔再次进行单独评定，即为“单独评定”水平。

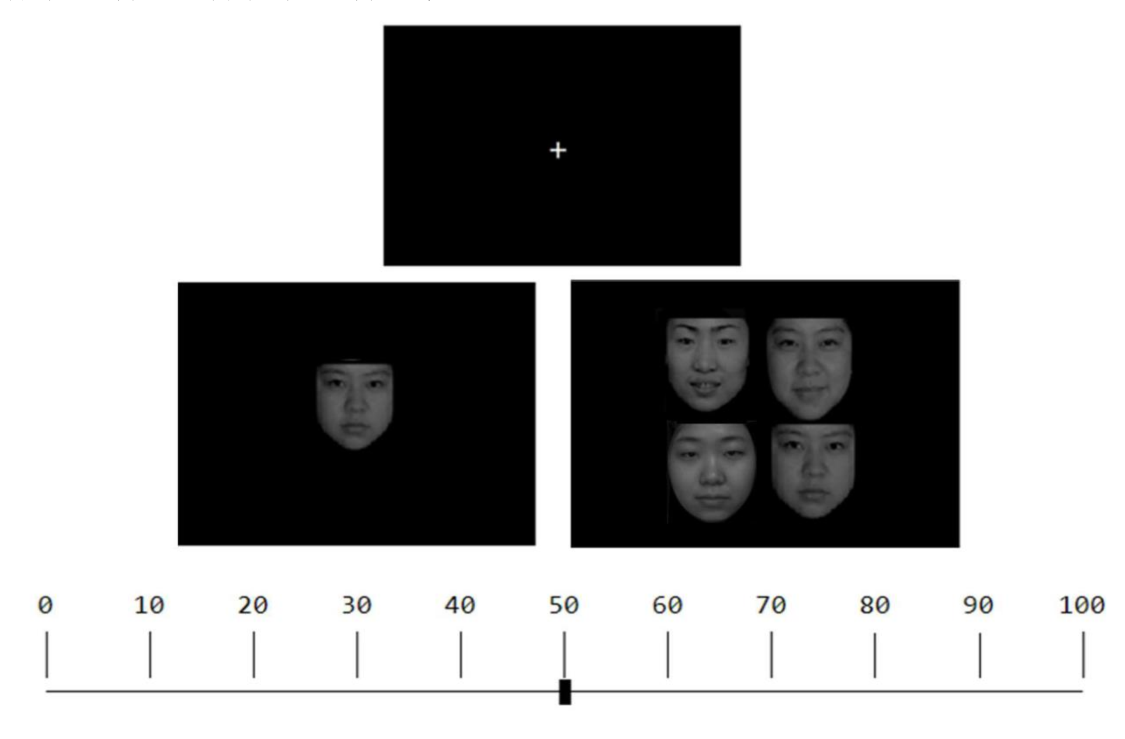


图 5　实验 3、4 流程图

（3）实验设计

采用单因素 5 水平（评分类型：不包含平均面孔的集合的成员均值 *M*1、不包含物理平均面孔集合但将平均面孔计算在内的成员均值 *M*2、不包含平均面孔的集合 G1、包含平均面孔的集合 G2、平均面孔 Avg）的被试内设计。因变量为被试对目标集合或目标面孔的吸引力评分。

（4）实验程序

实验流程如图 5 所示。首先呈现 500ms 注视点，而后在屏幕上呈现一组面孔或是一张单独面孔，被试要对目标的吸引力进行 0—100 的评分，0 代表吸引力最低，100 代表吸引力最高。

4.2 结果

由于单张面孔的评分任务中包含原始面孔也包含各个集合的平均面孔，因此我们随后使用单张面孔的评分来计算集合的吸引力平均值，计算不包含平均面孔集合条件下的集合成员评分均值，得到 *M*1 = 47.31；再假设该集合生成了平均面孔从而计算包含了平均面孔的成员均值，得到 *M*2 = 48.78。将 *M*1、*M*2、不包含

平均面孔条件的集合吸引力 G1、包含平均面孔条件的集合吸引力 G2 和平均面孔吸引力 Avg 作为评分类型 5 个水平进行方差分析。结果表明，评分类型主效应显著，F (4，112）= 27.60，$p < 0.001$，$\eta_p^2 = 0.50$。多重比较结果如下（如图 6)：

首先，M2 显著大于 M1，$p < 0.001$，95%CI = [1.22，1.71]，再次确认了合成平均面孔对集合吸引力平均值的提升作用。其次，不包含平均面孔的集合吸引力评分 G1 与包含平均的集合 G2 差异不显著，$p = 0.532$；与 M2 差异不显著，$p = 0.053$；但大于 M1，$p = 0.011$，95%CI = [1.26，8.80]。第三，平均面孔吸引力显著高于整个集合的吸引力 G1、G2 和成员平均值 M1、M2，p's <= 0.001。趋势分析表明，从集合成员平均值、集合吸引力到平均面孔，存在着逐渐增加的趋势，F (1，28）= 62.82，$p < 0.001$，$\eta_p^2 = 0.69$。

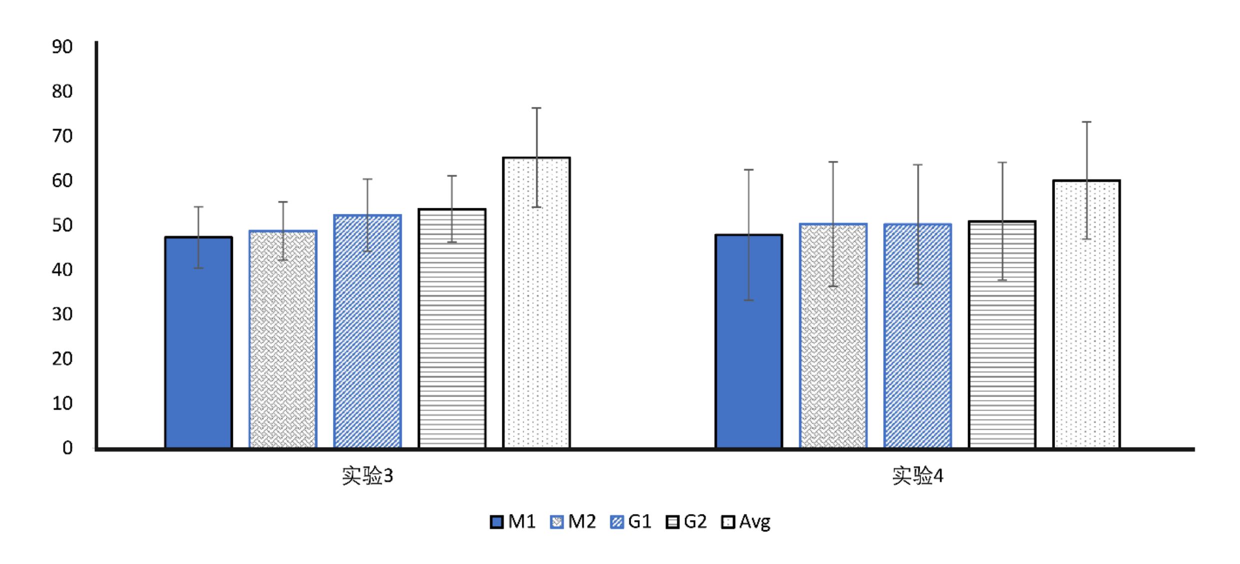


图 6 实验 3、4 吸引力评分结果（注：M1 为不包含平均面孔的集合的成员均值、M2 为不包含物理平均面孔集合但将平均面孔计算在内的成员均值、G1 为不包含平均面孔的集合、G2 为包含平均面孔的集合、Avg 为平均面孔）

除此以外，我们也尝试了分析平均面孔和集合吸引力的差值，发现在包含或不包含平均面孔的集合间差值没有显著差异，$t(28) = 0.19$，$p = 0.852$，再次验证了实验 1 中被试选择探测刺激平均面孔吸引力更高的比例在集合包含和不包含平均面孔条件下没有显著差异。

4. 3 讨论

实验 3 的评分任务结果基本重复了实验 1 的结果。首先，在大容量集合中的评分任务验证了大容量集合的集合吸引力高评现象，即集合吸引力高于集合成员

评分的平均值。其次，平均面孔吸引力大于集合吸引力，说明了实验 1 中被试判断探测面孔吸引力更高的原因。第三，集合是否包含平均面孔，对于集合评分没有显著影响（支持假设 2，不支持假设 1）；并且，不包含平均刺激的集合评分 G1 只有在考虑了生成平均面孔条件下才接近集合成员平均值，即 *M*2。最后，趋势分析和多重比较结果表明，不包含平均刺激的集合评分更接近于包含平均面孔条件的结果。这些结果说明大容量集合确实生成了平均面孔。

5 实验4：小容量面孔集合的评分

实验4采用小容量面孔集合进行评分任务，为不同容量下集合吸引力和平均吸引力的关系提供更直观的证据。

5.1 方法

（1）被试

采用 GPower 以统计功效 power=0.8，中等效应量 *f*=0.25 和单因素 5 水平(评分类型：不包含平均面孔的集合的成员均值 *M*1、不包含物理平均面孔集合但将平均面孔计算在内的成员均值 *M*2、不包含平均面孔的集合 G1、包含平均面孔的集合 G2、平均面孔 Avg)为参数估计的最小样本量为 *N*=21。实际招募中国人民大学在校生 31 名大学生，剔除一名评分全距小于 10 的被试，有效被试 30 名(15 名女性)，平均年龄 21.39 岁，标准差 2.46，右利手，视力或矫正视力正常。

（2）实验材料

实验刺激与实验 1 实验 2 相同。

集合刺激包含 4 张图片，集合刺激以 2×2 矩阵呈现。当评价类型为集合平均面孔，刺激呈现在屏幕中央，单张面孔图片的视角为 5.69° ×6.53° 。

在评分类型上，使用 4 张原始面孔组成集合，即是“不包含平均面孔集合”水平；如果使用 3 张原始面孔组成集合，并将集合成员的平均面孔作为新成员进入集合中，即是“包含平均面孔集合”水平，将集合成员面孔和平均面孔再次进行单独评定，即为“单独评定”水平。

（3）实验设计与程序

与实验 3 相同。

5.2 结果

与实验 3 类似，使用单张面孔的评分来计算集合的吸引力平均值，计算不包含平均面孔集合条件下的集合成员评分均值，得到 $M1 = 47.87$；再假设该集合生成了平均面孔从而计算包含了平均面孔的成员均值，得到 $M2 = 50.32$。将 $M1$、$M2$、不包含平均面孔条件的集合吸引力 G1、包含平均面孔条件的集合吸引力 G2 和平均面孔吸引力 Avg 作为评分类型 5 个水平进行方差分析。结果表明（如图 6），评分类型主效应显著，$F(4，116) = 6.27$，$p < 0.001$，$\eta_p^2 = 0.18$。多重比较结果如下：

首先，$M2$ 显著大于 $M1$，$p < 0.001$，95%CI = [1.82，3.08]，再次确认了合成平均面孔对集合吸引力平均值的提升作用。其次，不包含平均面孔的集合吸引力评分 G1 与包含平均的集合 G2 差异不显著，$p = 0.110$；与 $M2$ 差异不显著，$p = 0.977$，与 $M1$ 的差异也不显著，$p = 0.504$。第三，平均面孔吸引力显著高于整个集合的吸引力 G1、G2 和成员平均值 $M1$、$M2$，p's $\leq$ 0.007。此外，趋势分析表明，从集合成员平均值、集合吸引力到平均面孔，存在着逐渐增加的趋势，$F(1，29) = 21.05$，$p < 0.001$，$\eta_p^2 = 0.42$。

除此以外，集合平均面孔和集合整体吸引力的差值，不包含平均面孔集合条件大于包含平均面孔集合的条件(9.90 vs. 3.64)，$t(29) = 6.40$，$p < 0.001$，95%CI = [4.26，8.26]，Cohen's $d = 2.38$，可以更直观地印证实验 2 的包含平均面孔条件下中“被试判断探测刺激吸引力更高的比例降低了”这一结果。

对实验 4 和实验 3 中的平均面孔吸引力进行比较，结果表明小集合面孔形成的平均面孔吸引力更低，60.11 vs. 65.24，$p = 0.004$，95%CI = [2.76，13.85]，Cohen's $d = 0.94$。对实验 4 和实验 3 中平均面孔和集合平均值的差值(6.76 vs. 12.55)进行跨实验比较，发现存在一种可能的趋势，即实验 4 的小集合中平均面孔和集合成员平均值的差异更小，校正 $t(35.649) = 1.72$，$p = 0.094$，95%CI = [–1.06，12.81]，Cohen's $d = 0.07$。

5.3 讨论

实验 4 的评分任务结果为实验 2 的结果提供了直接的支持。首先，在小容量集合中，集合吸引力高评现象减弱了，集合吸引力和集合成员评分平均值 $M1$,$M2$ 都没有显著差异，结合实验 3 结果，验证了前人结论：集合吸引力高评现象在大容量集合强，在小容量集合弱。这符合平均面孔在小集合中更易受干扰的假设。

此外，小容量集合的平均面孔吸引力确实下降，和集合平均值的差异更小，因而小集合平均面孔吸引力相对大集合而言较低也是高评现象减少的一个可能原因。

其次，实验 4 中，类似于实验 3 的结果，集合是否包含平均面孔对于集合评分没有显著影响（支持假设 2，不支持假设 1）；并且不包含平均刺激的集合评分 G1 也和平均面孔计算在内的集合成员平均值 *M*2 没有差异。趋势分析和多重比较结果表明，不包含平均刺激的集合评分更接近于包含平均面孔条件的结果。这些结果说明小容量集合也可能受到了生成的平均面孔的影响。此外，集合吸引力和平均面孔吸引力的差异在不包含平均面孔刺激的集合条件下更大（假设 3），这反映了包含平均面孔刺激的集合吸引力更接近平均面孔吸引力，更直观地说明了在较小容量的集合中，被试主观形成的平均面孔表征被抑制（结合实验 2 结果），也说明了为何在实验 2 的包含平均面孔条件下中，被试判断探测刺激吸引力更高的比例降低了。

6 综合讨论

实验 1、2 的按键反应和扩散模型拟合结果和实验 3、4 的评分结果共同说明，当集合容量为 12 张，高吸引力的平均面孔是否出现不影响集合吸引力评分和平均辨别任务，说明表征集合时形成了集合平均面孔；当集合容量为 4 张，平均面孔效应减弱了，可能是由于集合平均面孔的吸引力较低以及平均面孔被个体表征干扰了。

6.1 集合面孔吸引力高评现象

在 Van Osch 等人(2015)的研究中，通过采用给若干名女性拍摄的自然材料(如聚会照片)评分来探究集合吸引力，则发现了在较大的集合(人数较多的照片)中存在集合吸引力高于集合成员吸引力平均值的情况。

实验 1 的平均辨别任务发现，当集合不包含平均面孔时，判断探测刺激吸引力更高的比例与包含平均面孔时相似，说明两种条件的平均吸引力相近，不包含平均面孔的集合吸引力高于成员面孔的平均值，从另一个角度反映了集合吸引力高评现象。实验 3 则直接再次验证了大容量集合的集合吸引力高评现象。

实验 4 的评分结果说明集合吸引力高评现象较弱。而实验 2 正确率分析表明，实验 2 也出现了集合吸引力高评现象，只是实验 1 更明显。类似地，在 Van Osch 等人(2015)的研究中，较小的集合也很少观察到集合吸引力高于成员吸引力平均

值。在本研究中，既有直接通过评分获得集合面孔吸引力高评现象的直接证据，也有借助平均面孔在集合中的作用大小来推测的部分。Van Osch 等人(2015)研究中使用的是生态效度较高的自然材料，存在成员吸引力分布集中，缺少具有代表性的高低吸引力面孔的问题。本研究则改用了没有背景，由单独评分的面孔素材组成的集合，并平衡不同吸引力水平的面孔数量，依旧得到了类似的结论，说明在大容量集合中集合吸引力高评现象是比较稳定的，且现象的产生与集合平均面孔的形成有关。

6.2 平均表征的形成机制

通过将平均辨别的实际反应与不同理论假设的反应分布进行比较和拟合，我们发现集合吸引力包含平均面孔贡献的假设拟合更好，支持了平均表征过程中形成了平均刺激(如平均面孔)。在实验 1 中，集合包含平均面孔时，判断平均面孔吸引力更高的比例并未降低，说明包含或不包含平均面孔刺激的集合吸引力同等接近平均面孔，因此平均表征并不是通过集合成员的平均值计算得到的。类似的，无论还是实验 3 还是实验 4，都发现平均表征的吸引力要远高于集合成员的平均值。在实验 2 的小集合中，发现新面孔作为探测刺激与平均面孔作为探测刺激的结果不同，则说明平均面孔作为探测刺激出现之前已经得到了加工。这个结论和近期 Ying 等人（2020）的结论一致，他们通过吸引力适应后效范式发现，由一组面孔引发的适应后效等于这一组的平均面孔引发的适应后效，同样支持了集合表征中存在平均刺激的形成。

值得注意的是，平均刺激的形成也需要资源投入，体现在实验 1 没有平均面孔输入时需要更长的编码加工时间。Huang(2015)发现对于对象特征和统计表征，启动的效果是相等的，说明统计表征的形成至少需要和单个个体加工同等的注意资源。因此，平均表征和个体表征在早期可能是因此相互竞争的关系(Li et al，2016)。由于早期没有足够的认知资源加工所有个体，因此优先形成了平均表征。Bauer (2017)以不同长短的线条作为集合刺激，在平均辨别任务前加入数字记忆任务的研究发现，相比低记忆负荷条件(1 个 0)，高记忆负荷条件下(4~7 个随机数字)更有利于形成平均表征。实验 1 和实验 2 的跨实验比较也发现，在小集合中整体信息积累速度也比大集合更慢。

与此同时，小集合同样形成了平均表征且需要资源的投入，表现在实验 2

没有平均面孔输入时信息积累较慢。也就是说无论大小集合，都有平均刺激的形成。而大小集合之间的反应差异，是由于其他原因造成的，而不是出于不同的加工机制。

6.3 集合吸引力与平均面孔的关系

尽管集合吸引力表征中包含了平均面孔的贡献，但集合吸引力并不完全等同于平均面孔的吸引力。无论实验 1 还是实验 2，均发现被试倾向于评价平均面孔比集合吸引力更高的结果。并且实验 3、4 平均面孔的评分也要高于集合的评分，说明集合吸引力是基于平均面孔的形成，将平均刺激纳入进来成为集合的成员，再对于集合进行整体评价。因此，集合吸引力高评现象并不完全如 Van Osch 等人(2015)所推断的仅依赖于平均面孔，而是与表情集合加工类似，读取平均面孔的表征(Haberman & Whitney，2009)。

类似的现象是，被试经常在再认任务中把平均表征误以为是集合中的成员，或者认为平均面孔与其中一个成员具有相同的身份(Neumann et al.，2013)。判断过程中被试可能使用了一种策略：在加工整个集合吸引力的时候，将平均面孔知觉成集合中某一成员，由于观察到集合中的其他面孔的吸引力多数低于平均面孔，进而倾向于认为单独出现的平均面孔的吸引力较高。

研究者曾发现，当存在多张面孔，单张面孔的吸引力会被判断得比孤立出现时更高(Walker & Vul，2014)，并同样解释为平均的作用。背景面孔导致会将人脸的感知偏向群体的平均水平。因此，伴随同一目标面孔出现的背景面孔吸引力越高，目标面孔就会被评价得越高（DeBruine et al.，2007；Perrett et al.，1994；Walker & Vul，2014）。因此，平均面孔产生可能会对集合中的其他成员产生影响，由于平均面孔的吸引力高，导致集合其他成员的吸引力也得到了抬升。这是平均面孔对其他面孔产生的影响，也可能是集合吸引力高评现象产生的路径之一。

6.4 平均刺激表征与集合容量的关系

小容量集合中面孔吸引力一般会被评价与集合成员平均值相同(Anderson，1965; Anderson et al.，1973)，且集合吸引力高评现象在小容量集合消失(Van Osch et al.，2015)，说明小容量的平均表征存在一些特异性。

一方面，小容量集合产生的平均面孔刺激的吸引力相对较低，可能因此导致对于集合吸引力的增加效果较小，因此集合吸引力高评现象减少。另一方面，可

能是平均表征收到了干扰。逆层级理论提出(Hochstein et al.，2015)，高级皮层的整体表征以自上而下的方式返回到局部加工，以局部细节信息证实(或矫正)初步的整体表征估计值，也就是说平均表征在加工后期会受到个体表征的矫正。Li et al(2016)也发现，当呈现时间延长，认知资源增加，会导致个体表征的精度增强。

这两种理论均得到了本研究结果的支持，在实验2中判断平均面孔吸引力更高的比例更低，以及实验4中的平均面孔和集合吸引力差值比较结果表明小集合包含平均面孔的集合吸引力和平均面孔的差异更小，都说明小集合平均面孔吸引力较低在其中存在影响。与此同时，实验2在包含平均面孔刺激时平均面孔吸引力更高的比例更低，说明相对于被试自己生成平均面孔刺激，直接输入平均面孔减少了集合吸引力和平均面孔之间的差异，支持小集合的平均面孔受到了干扰。在小集合中无论集合是否包含平均面孔，信息积累速度都比大集合更慢，也支持逆层级理论所说的局部细节信息对于整体表征的矫正和干扰。

7 结论

(1)面孔集合吸引力高评现象是基于平均面孔刺激的形成。

(2)平均表征的产生是基于平均刺激的产生。

(3)无论在大小集合中都形成了平均刺激。小集合中集合吸引力不出现高评现象是由于平均刺激受到了干扰，并且平均刺激本身吸引力较低。

(4)平均表征的加工也需要一定认知资源。

参考文献:

Abbas，Z. A.，& Duchaine，B. (2008). The role of holistic processing in judgments of facial attractiveness. *Perception，37*，1187－1196.

Alvarez，G. A.，& Oliva，A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science，19*(4)，392－398.

Alvarez，G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences，15*(3)，122－131.

Anderson，N. H. (1965). Averaging versus adding as a stimulus combination rule in impression formation. *Journal of Experimental Psychology*，*70*，394－400.

Anderson，N. H.，Lindner，R.，& Lopes，L. L. (1973). Integration theory applied to judgments of group attractiveness. *Journal of Personality and Social Psychology，26*，400－408.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157 - 162.

Bauer, B. (2009). Does Stevens' s power law for brightness extend to perceptual brightness averaging? *Psychological Record, 59*(2), 171 - 185.

Bauer, B. (2017). Perceptual averaging of line length: Effects of concurrent digit memory load. *Attention, perception & psychophysics, 79*(8), 2510 - 2522.

Brady, T. F., & Alvarez, G. A. (2015). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 921 - 929.

Carragher, D. J. , Lawrence, B. J. , Thomas, N. A. , & Nicholls, M. E. R. (2018). Visuospatial asymmetries do not modulate the cheerleader effect. *Scientific Reports, 8*(1), 2548.

de Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics, 70*(5), 789 - 794.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), R751 - R753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology, 35*, 718 - 734.

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe & L. Robertson (Eds.). *Oxford series in visual cognition. From perception to consciousness: Searching with Anne Treisman* (p. 339 - 349). Oxford University Press.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General, 144*(2), 432 - 446.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36*(5), 791 - 804.

Hochstein, S., Pavlovskaya, M., Bonneh, Y. S., & Soroker, N. (2015). Global statistics are not neglected. *Journal of Vision, 15*(4), 7.

Huang, L. (2015). Statistical properties demand as much attention as object features. *Plos One, 10*(8), e0131191.

Ji, L., Chen, W., Loeys, T., & Pourtois, G. (2018). Ensemble representation for multiple facial expressions: Evidence for a capacity limited perceptual process. *Journal of Vision, 18*(3), 17, 1–19.

Komori, M., Kawamura, S., & Ishihara, S. (2009). Averageness or symmetry: Which is more important for facial attractiveness? *Acta Psychologica, 131*, 136–142.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science, 1*(2), 115–121.

Li, H., Ji, L., Tong, K., Ren, N., Chen, W., Liu, C. H., Fu, X. (2016). Processing of individual items during ensemble coding of facial expressions. *Frontiers in Psychology, 7*:1332.

Luo, A. X., & Zhou, G. (2018). Ensemble perception of facial attractiveness. *Journal of Vision, 18*(8):7, 1–19.

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision, 15*(4).

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics, 70*(5), 772–788.

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition, 128*(1), 56–63.

O' Toole, A. J., Price, T., Vetter, T., Bartlett, J. C., & Blanz, V. (1999). 3D shape and 2D surface textures of human faces: The role of "averages" in attractiveness and age. *Image and Vision Computing, 18*, 9–19.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739–744.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two–choice decision tasks. *Neural Computation, 20*(4), 873–922.

Rhodes, G. , Tremewan, T. (1996). Averageness, exaggeration, and facial attractiveness. *Psychological Science, 7*, 105 – 110

Rhodes, G., Yoshikawa, S., Clark, A., Lee, K., McKay, R., & Akamatsu, S. (2001). Attractiveness of facial averageness and symmetry in nonwestern cultures: In search of biologically based

standards of beauty. *Perception, 30*, 611–625.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two–choice response times. *Psychological Methods, 16*(1), 44–62.

Van Osch, Y., Blanken, I., Meijs, M. H. J., & Van Wolferen, J. (2015). A group's physical attractiveness is greater than the average attractiveness of its members: the group attractiveness effect. *Personality and Social Psychology Bulletin, 41*(4), 559–574.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner P. (2019). Rank–normalization, folding, and localization: An improved R–hat for assessing convergence of MCMC. *Bayesian Analysis*, Advance publication (2021), 28 pages.

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology, 60*(6), 385.

Walker, D., & Vul, E., (2014). Hierarchical encoding makes individuals in a group seem more attractive. *Psychological Science, 25*(1), 230–235.

Wang, Y., & Luo, Y. J. (2005). Standardization and assessment of college students' facial expression of emotion. *Chinese Journal of Clinical Psychology, 13(4)*, 396–398.

[王妍, 罗跃嘉. (2005). 大学生面孔表情材料的标准化及其评定. *中国临床心理学杂志, 13*(4), 396–398.]

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology, 69*(1), 105–129.

Willis, R. H. (1960). Stimulus pooling and social perception. *Journal of Abnormal and Social Psychology, 60*, 365–373.

Ying, H., Burns, E., Choo, A. M., & Xu, H. (2020). Temporal and spatial ensemble statistics are formed by distinct mechanisms. *Cognition, 195.*

作者贡献声明：

陈文锋：提出研究问题及框架，设计研究方案，分析数据，论文最终版本修订；

田欣然，侯文霞：进行实验，分析数据，论文起草；

欧玉晓，易冰：采集数据；

尚俊辰：论文最终版本修订；